**Abstract**

There are many varieties of Avocado in the U.S. markets however this project will analyze data from the Hass avocado variety. Having important information on current and future prices and sales volume of avocado is key to sales forecasting, adequate planning, risk management, increase revenue, precise decision-making and innovation. To accomplish this task, two Predictive analytics models will be used as the best assessment tool in predicting future prices of avocado. In this study, predictions on average prices of avocado will be made given: a region, a year, a type (conventional or organic) and a week for a particular year. Sales volume predictions of avocado will be made given: year, type and week. For instance, in the region of Albany, year 2015, second week of January, can or will my model predict the total volume of organic avocado as 1676.05? A total of 18,250 records and 15 attributes will be analyzed. Some of the key attributes are Date, Average Price, Type, Region, Total volumes etc. These data collected from various outlets reporting consists of a combination of the channels such as: grocery, mass, club, drug, dollar and military. They can be found in the Hass Avocado Board website. Furthermore, these two models will be analyzed, trained, evaluated and compared. The model with the highest prediction accuracy and lowest error rate will be recommended. The prediction models to be used are Multiple Linear and Multiple Non-Linear Regression. Machine learning algorithms will be explored as well etc. R programming will be the main tool for statistical analysis and tableau might be used for some form of visualization.

**Literature Review:**

**Introduction**

Avocados are economically valuable fruits and are very popular in the U.S. and worldwide. They are highly nutritious and contain a wide range of vitamins and minerals. They have so many health benefits. For example, they are powerful antioxidants, loaded with fiber, easy to incorporate in diets, help prevent cancer, weight loss and have healthy monounsaturated fatty acids which is good for the heart. They are mainly grown in Mexico, Central and South American Countries like Peru and Chile. They are readily available all year round in the U.S. even though Avocados are grown seasonally in the United States.

In 2002, the Hass Avocado Board(HAB) was set-up primarily to promote Hass Avocado in the United States. Its website [www.HassAvocadoBoard.com](http://www.HassAvocadoBoard.com) serves as an online resource center which furnishes relevant and up-to-date data for local producers and importers of Avocado in the United States. This information includes shipment data, retail data and research, and consumer research.

**1**In 2017, the U.S. produced 146, 310 tons of avocado valued at $392 million with California accounting for about 88.2% of avocado production followed by Florida (11.5%) and Hawaii (0.3%) respectively. The demand, consumption rate and price for avocados in the U.S. has increased over the years. That notwithstanding, several factors affect the price and sales volume of avocados in the U.S. Avocado does not thrive during the winter seasons which leads to decrease in supply or production and increase in price. This causes U.S. to import most of its avocados from Mexico and other foreign growers to maintain supply. Trends in economic and political atmosphere in the contributing countries affect the price of avocado in the U.S. For instance, the recent NAFTA trade pact crises now known as USMCA (United States Mexico Canada Agreement) lead to increase in price. Some disgruntled growers in some of the contributing countries are heading towards Mexico and the U.S. in search of greener pastures due to crises in their countries of origin thus leading to reduction in avocado production in these countries which as a result leads to price increase. **2**As at August 2018, Avocado price rose rapidly in the U.S. and is expected to continue for another 9 to 12 weeks. This rapid increase has caused Avocado to be sold for double the price abruptly in North America.

The aim of this research is to analyze current outcomes and make predictions about future outcomes through the use of historical data, a variety of statistical techniques from predictive modelling, machine learning and data mining. These predictions will serve as the best assessment tool for future trends of Avocados in the United States as well as solving complex problems, improving market campaign and operations, making key decisions, increase profits, providing valuable insight of the avocado industry, discovering innovative prospects and to help the United States to be well prepared for events that cause rapid fluctuations in the price and sales volume of avocado.

As stated earlier, Multiple Linear and Multiple Non-Linear Regression Models will be the main technique used. Other techniques may be considered if necessary. While searching articles for my literature review, it was observed that there were only a handful of articles where multiple linear regression and multiple non-linear regression were used as prediction pricing models.

**1**[www.statista.com/statistics/610460/production-avocados-us-by-state](http://www.statista.com/statistics/610460/production-avocados-us-by-state)

**1**[www.agmrc.org/commodities-products/fruits/avocados](http://www.agmrc.org/commodities-products/fruits/avocados)

**2**[www.freshfruitportal.com/news/2018/08/10/u-s-avocado-market-prices-rise-sharply](http://www.freshfruitportal.com/news/2018/08/10/u-s-avocado-market-prices-rise-sharply) -

(Source: USDA Market News via [Agronometrics](http://www.agronometrics.com/" \t "_blank))

**Literature Review:**

Three articles were reviewed to illustrate how the multiple linear regression models are used as prediction models.

1In this article, the authors explored the use of multiple linear regression to predict rice cultivation time based on the weather analysis and to escalate Farmer’s Exchange Rate. Weather data from National Weather Forecast and Farmer's Exchange Rate data from National Statistics Authority for the past one year. Then, the data was cleaned by removing unwanted data thus improving the quality of data collected. Further pre-processing included outliers detection, selection of the required data type, best features was selected and the data trained based on the season. Being a multiple linear regression model, the correlation between the dependent variable and multiple independent variables was easily determined. In this case, Farmer’s Exchange Rate denoted by (yi) is the dependent variable while the multiple independent variables are “Avg. Temperature(x1)”, “Avg. Humidity(x2)”, “Rainfall(x3)” and “Solar Radiation(x4)”. Determining the regression coefficients was crucial in constructing the model:

yi  = β0 + β1x1 + β2x2 + β3x3 + β4x4, β0 is the constant.

Root Mean Square Error was used in testing the performance or the prediction accuracy rate of the model by comparing the predicted results to the actual records. The authors concluded that the model is a good and more result-driven model for predicting rice cultivation time even when compared with other machine learning models.

# 2In this article, the authors investigated the use of Multiple Linear regression model in predicting the annual energy consumption in the Spanish banking sector. Three regression models were developed for the prediction. The dataset was cleaned, and some independent variables that were not significant in the regression model were removed. Correlation was established between dependent and multiple independent variables. Least square method was used in obtaining the regression model. Outliers was taken into consideration because it can affect the outcome of the regression model after which Minitab tools “Best Subsets” and “Stepwise” Minitab tools were used in choosing the best model amongst the three models. The authors concluded that model 1 is the best model because it had the lowest determination coefficient. Thus, it is a reliable model for predicting the energy consumption in the Spanish banking sector.

# 3In this article, the authors studied the use of Linear Regression and Non-Linear Regression models in predicting surface wind components based on their Transfer Function(TF). Three Non-Linear regression models were considered namely Neural Network(NN), support vector machines(SVM) and Random Forest(RF).

# Problem of overfitting became more obvious when the number of observations used was reduced. When the predictability of the linear and non-linear regression models was compared, the non-linear regression model performed better than the linear model.

Several predictive analytics models were used, especially the regression models, in the above reviewed articles. The key is to build a model where errors can be reduced to the barest minimum in other to select the best model. Reduction in the number of variables selected helped in minimizing errors. Caution should be applied when fitting data into the model because overfitting the model will cause high variance and low bias while underfitting the model causes low variance and high bias in the model. To avoid overfitting the model, data was transformed, and new ones extracted, features were manually chosen. Relationship between the dependent and independent variables was easily established because it is a regression model. Data was preprocessed to obtain a high-quality dataset for analysis. Significant variables were selected, and regression coefficients were optimized to obtain higher prediction accuracy rates. Error measurement was used inn model evaluation. Some of these processes will be applied in my study. For instance, data will be preprocessed and model trained by using cross-validation to avoid overfitting.

In conclusion, multiple linear regression performed better when compared with other machine learning models like the Artificial Neural Network models; because of its ease of understanding, feasibility, higher efficiency, more result -driven and practical model.

**1Luminto, Harlili** “Weather analysis to predict rice cultivation time using multiple linear regression to escalate farmer's exchange rate”

# Published in: 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)

# Date of Conference: 16-18 Aug. 2017, Date Added to IEEE Xplore: 02 November 2017, ISBN Information:

# INSPEC Accession Number: 17334091, DOI: 10.1109/ICAICTA.2017.8090974, Publisher: IEEE

# 2Alfonso Aranda, Germán Ferreira, M.D.Mainar-Toledo, Sabina Scarpellini, Eva Llera Sastresa

"Multiple regression models to predict the annual energy consumption in the Spanish banking sector”

CIRCE – Centre of Research for Energy Resources and Consumption – University of Zaragoza, Mariano Esquillor Gómez, 15 – 50018 Zaragoza, Spain

Received 6 December 2011, Revised 7 February 2012, Accepted 24 February 2012, Available online 5 March 2012.

# 3Yiwen Mao, Adam Monahan

# “Linear and nonlinear regression prediction of surface wind components”

# Climate Dynamics, 2018, Volume 51, Number 9-10, Page 3291

**Dataset**

Dataset used in the project was downloaded from Kaggle website. It is in .csv file format and comprises of 18,250 records and 15 columns. Some of key attributes in the table are: Date, Average Price, Type, Year, Region and Total Volume. The data is sourced from multi-outlet reporting which includes a combination of the following channels such as grocery, mass, club, drug, dollar and military. In the table, the Average Price (of avocados) indicates a per unit (per avocado) cost, even when multiple units(avocados) are sold in bags. They are three Product Lookup codes(PLU’s) in the table: 4046, 4225 and 4770 for the Hass variety.

**Date** - The date of the observation

**AveragePrice** - the average price of a single avocado

**Type** - conventional or organic

**Year** - the year

**Region** - the city or region of the observation

**Total Volume** - Total number of avocados sold

**4046** - Total number of avocados with PLU 4046 sold

**4225** - Total number of avocados with PLU 4225 sold

**4770** - Total number of avocados with PLU 4770 sold

This is a link to the complete dataset and data dictionary:

<https://www.kaggle.com/neuromusic/avocado-prices/home>

**Approach:**

**Data Collection**

Load Avocado dataset from Kaggle website into R Studio for analysis

**Data Preparation**

Initial Analysis, Exploratory analysis, transformation, Dimensionality Reduction

**Experimental Design**

Split data into training and test sets, Treatment for imbalance, cross-validation

**Modeling**

Regression

**Evaluation**

Regression: R-Squared, goodness of fit, RMSE(root mean square error)

**Compare the two Models**

Work in Progress.

**Accept Model**

**Step 1: Data Collection:**  Load data from Kaggle website:

<https://www.kaggle.com/neuromusic/avocado-prices/home>

**Step 2: Data Preparation:** check my dataset for errors, inconsistencies, missing values, low variance, pairwise relations or correlation analysis, inconsistencies, outliers, 5 number Summary: max, min, mean, first quartile and third quartile). Check if normalization is needed, sub-setting data, multivariate relations, data reduction, visualization.

Sub-set data: the “date attribute” in other to analyze weekly price of Avocados. Investigate to make sure that multiple independent variables in the dataset are not correlated- Multicollinearity. If they are, remove the highly correlated independent variables

**Step 3: Experimental Design:** Data will be split into Training and test sets. Training data will be used for training the model and Test data will be used to validate the model.

Step 4: **Modelling:** Build Multi Linear Regression analysis model to predict the Avocado price.

.**Step 5: Evaluation:** Measure the accuracy of the model using test data.

Evaluate models by using error measurements such as R squared goodness of fit test for the Multiple Linear model and **Standard error of the Regression** and other testsfor the Multiple Non-Linear Regression model.

**Step 6: Compare Models**: Compare the two models and choose the model with the highest prediction accuracy rate. Is there any room for improvement of the chosen model? If the accuracy is too high, this could be a problem of overfitting.

**Step 7: Accept Model:** Accept model.

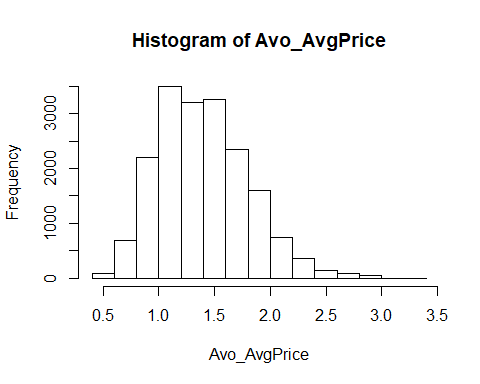
**Initial Results:**

**Column 2: AveragePrice**

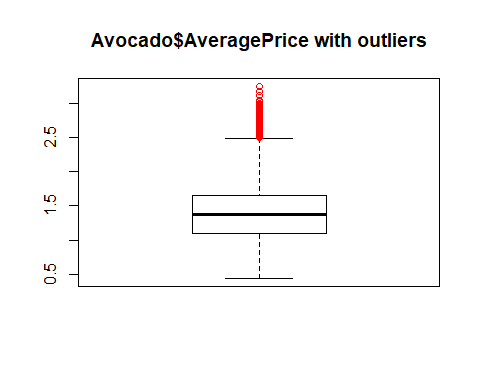
summary(Avo\_AvgPrice)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.440 1.100 1.370 1.406 1.660 3.250

hist(Avo\_AvgPrice, breaks = 20) ### ----> This goes to the intial analysis <---



boxplot(Avo\_AvgPrice, main = "Avocado$AveragePrice with outliers", horizontal = F, outcol = "red") ### ----> This goes to the INITIAL analysis <---

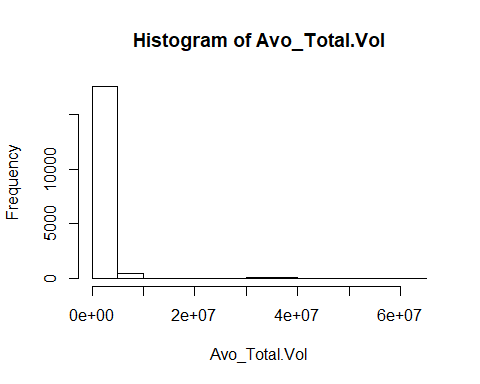


**Column 3: Total.Volume**

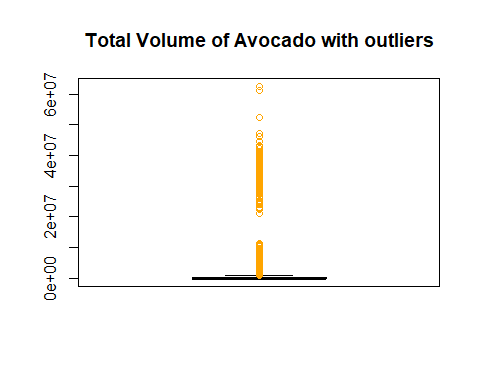
summary(Avo\_Total.Vol)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 85 10839 107377 850644 432962 62505647

hist(Avo\_Total.Vol) ### ----> This goes to the intial analysis <---



boxplot(Avo\_Total.Vol, main = "Total Volume of Avocado with outliers", horizontal = F, outcol = "orange") ### -> This goes to the intial analysis <-

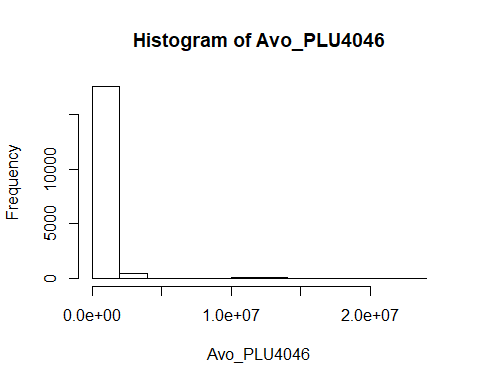


# Column 4: PLU4046

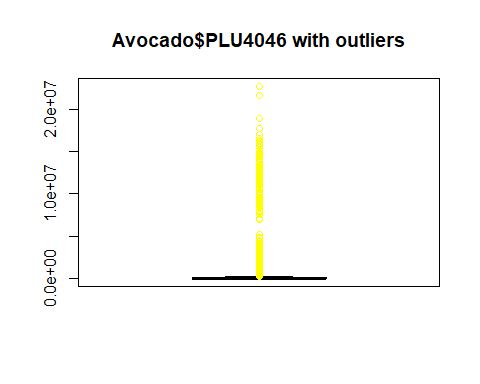
summary(Avo\_PLU4046)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 854 8645 293008 111020 22743616

hist(Avo\_PLU4046)



boxplot(Avo\_PLU4046, main = "Avocado$PLU4046 with outliers", horizontal = F, outcol="yellow")

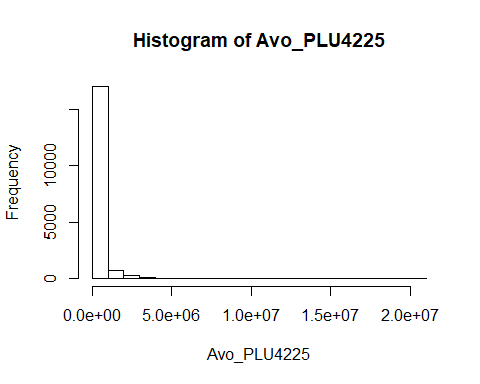


**Column 5: PLU4225**

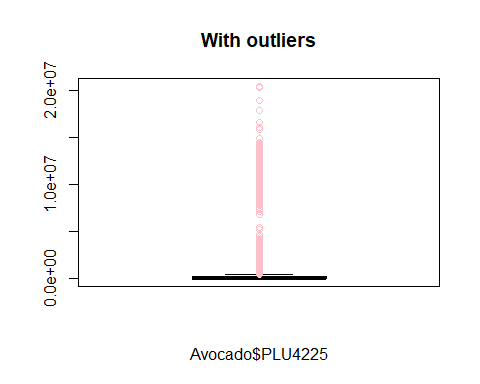
summary(Avo\_PLU4225)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 3009 29061 295155 150207 20470573

hist(Avo\_PLU4225)



boxplot(Avo\_PLU4225, main = "With outliers", xlab= "Avocado$PLU4225", horizontal = F, outcol="pink")

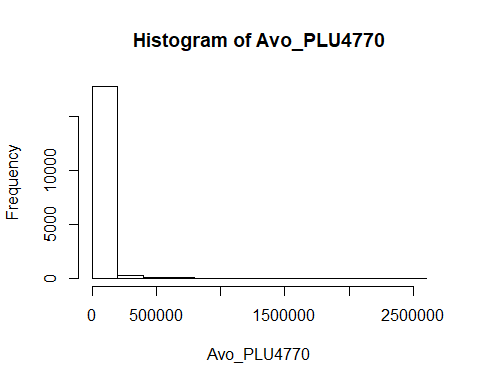


**Column 6: PLU4770**

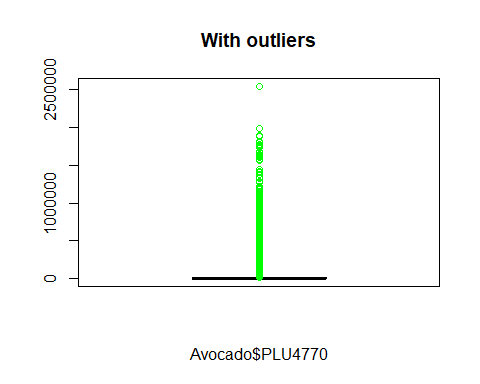
summary(Avo\_PLU4770)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 185 22840 6243 2546439

hist(Avo\_PLU4770)



boxplot(Avo\_PLU4770, main = "With outliers", xlab= "Avocado$PLU4770", horizontal = F, outcol="green")

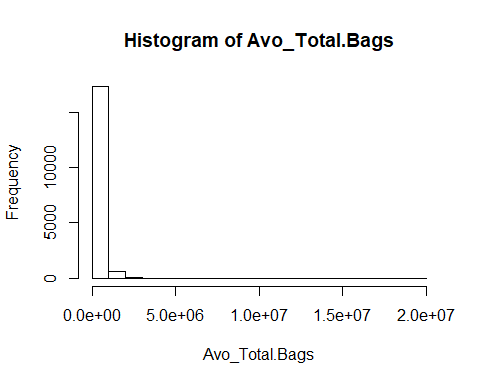


**Column 7: Total.Bags**

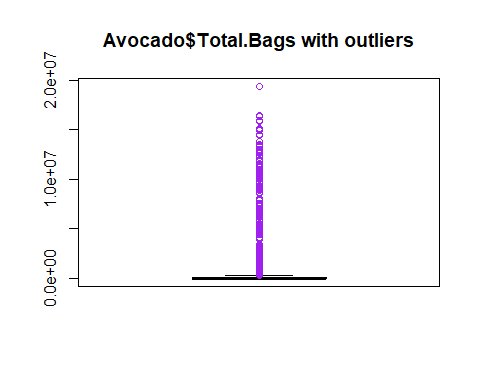
summary(Avo\_Total.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 5089 39744 239639 110783 19373134

hist(Avo\_Total.Bags)



boxplot(Avo\_Total.Bags, main = "Avocado$Total.Bags with outliers", horizontal = F, outcol = "purple")

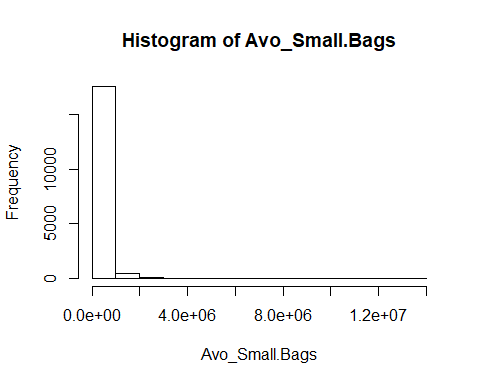


**Column 8: Small.Bags**

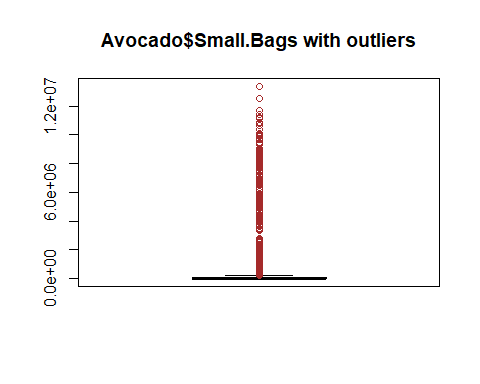
summary(Avo\_Small.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 2849 26363 182195 83338 13384587

hist(Avo\_Small.Bags)



boxplot(Avo\_Small.Bags, main = "Avocado$Small.Bags with outliers", horizontal = F, outcol = "brown")

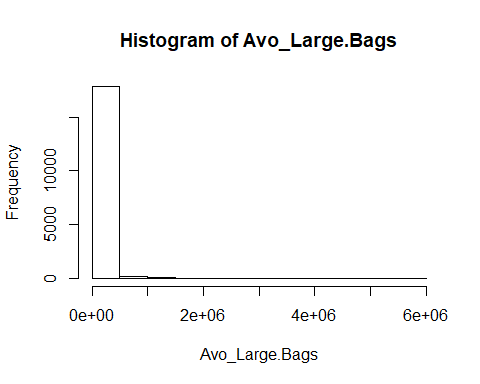


**Column 9: Large.Bags**

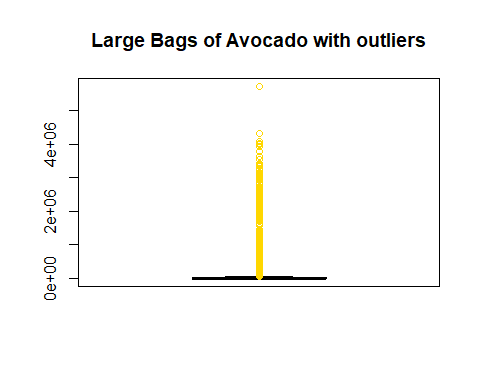
summary(Avo\_Large.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 127 2648 54338 22029 5719097

hist(Avo\_Large.Bags)



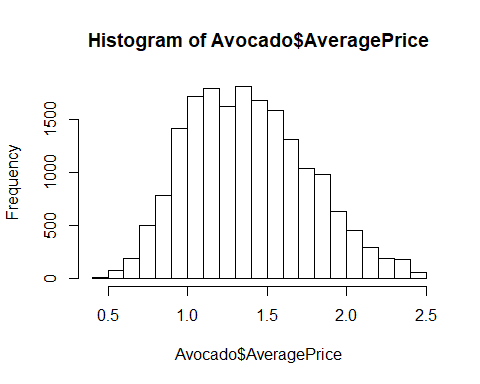
boxplot(Avo\_Large.Bags, main = "Large Bags of Avocado with outliers", horizontal = F, outcol = "gold")



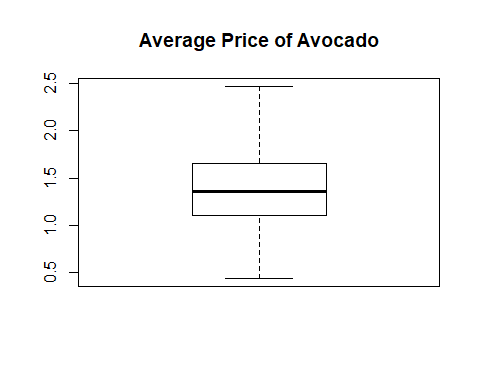
**Final Results:**

**Column 2: AveragePrice**

hist(Avocado$AveragePrice) ### ----> This goes to the FINAL analysis <---



boxplot(Avocado$AveragePrice, main ="Average Price of Avocado") ### ----> This goes to the FINAL analysis <---

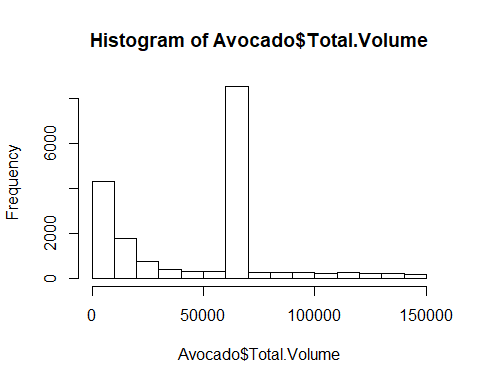


summary(Avo\_AvgPrice)

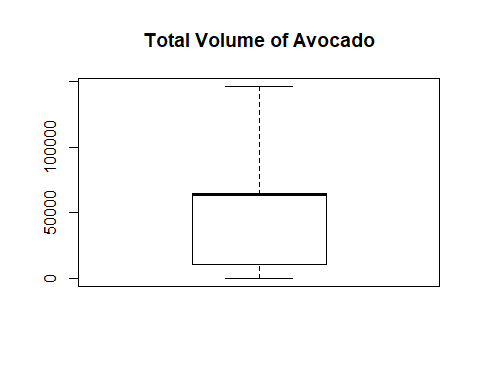
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.44 1.10 1.36 1.39 1.65 2.47

# Column 3: Total.Volume

hist(Avocado$Total.Volume) ### ----> This goes to the FINAL analysis <---



boxplot(Avocado$Total.Volume, main = "Total Volume of Avocado") ### ----> This goes to the FINAL analysis <---

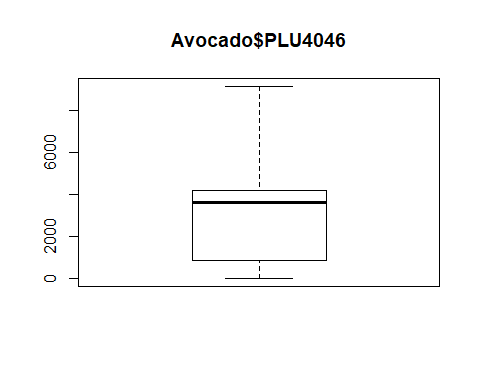


summary(Avo\_Total.Vol)

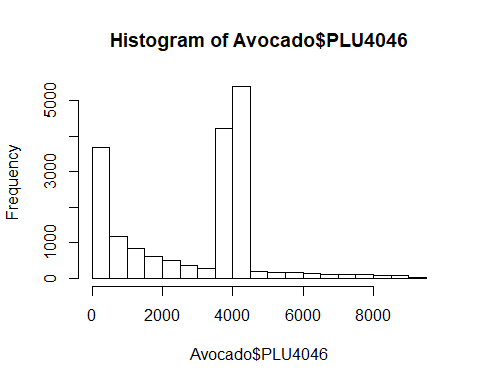
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 84.56 10838.58 63902.90 47243.99 65098.19 146422.18

# Column 4: PLU4046

boxplot(Avocado$PLU4046, main="Avocado$PLU4046")



hist(Avocado$PLU4046)

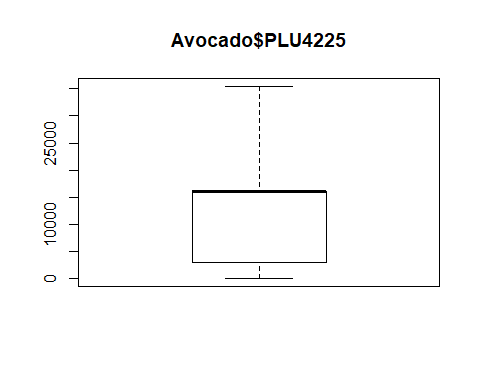


summary(Avo\_PLU4046)

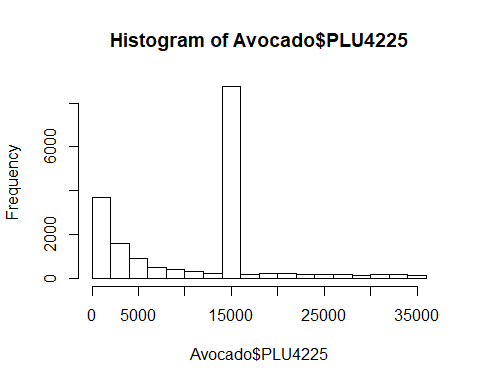
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 854.1 3606.8 2865.0 4163.6 9127.1

# Column 5: PLU4225

boxplot(Avocado$PLU4225, main="Avocado$PLU4225")



hist(Avocado$PLU4225)

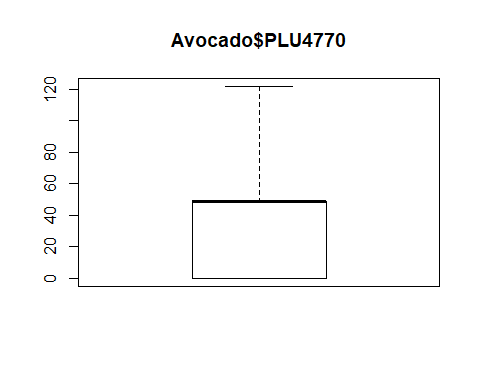


summary(Avo\_PLU4225)

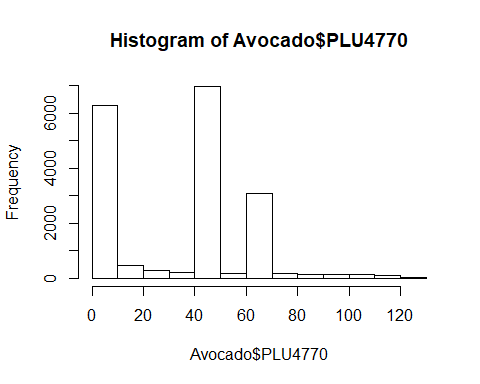
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 3009 15989 11550 15989 35457

# Column 6: PLU4770

boxplot(Avocado$PLU4770, main = "Avocado$PLU4770")



hist(Avocado$PLU4770)

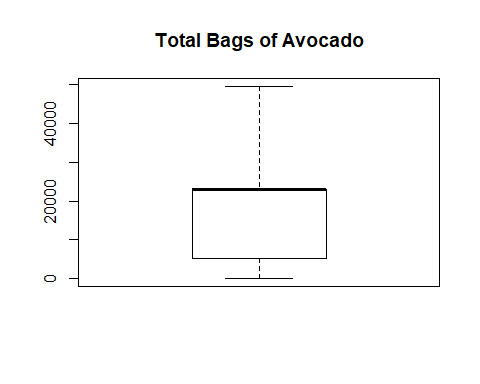


summary(Avo\_PLU4770)

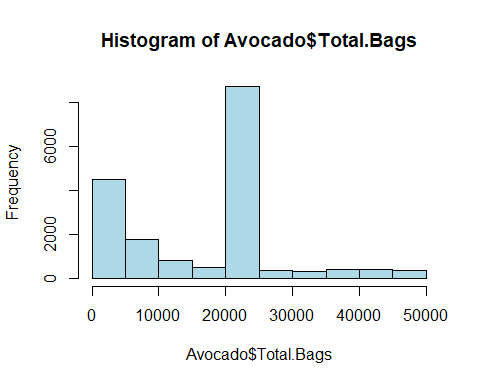
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 48.81 35.01 48.81 121.93

# Column 7: Total.Bag

boxplot(Avocado$Total.Bags, main="Total Bags of Avocado")



hist(Avocado$Total.Bags, col = "light blue")

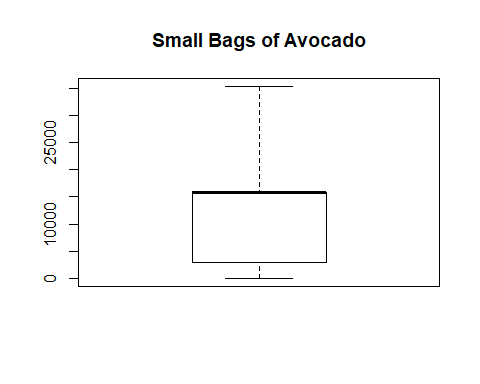


summary(Avo\_Total.Bags)

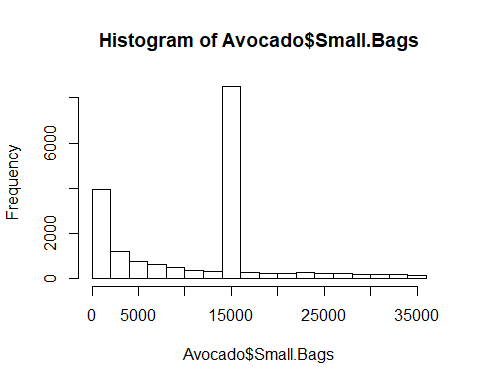
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 5089 22905 17054 22905 49622

# Column 8: Small.Bags

boxplot(Avocado$Small.Bags, main="Small Bags of Avocado")



hist(Avocado$Small.Bags)

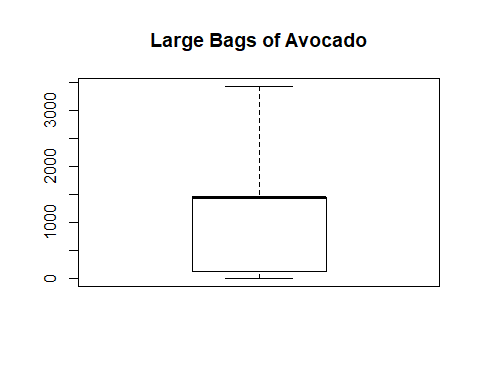


summary(Avo\_Small.Bags)

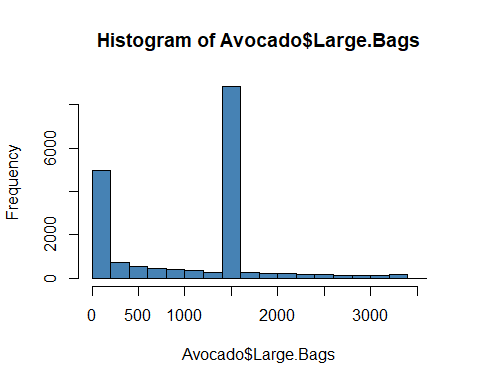
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 2849 15847 11645 15847 35322

# Column 9: Large.Bags

boxplot(Avocado$Large.Bags, main="Large Bags of Avocado")



hist(Avocado$Large.Bags, col = "steel blue")



summary(Avo\_Large.Bags)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 127.5 1446.0 1033.8 1446.0 3423.5